JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

# Category Query Learning for Human-Object Interaction Classification

Chi Xie[1], Fangao Zeng[2], Yue Hu[3], Shuang Liang[1]*, Yichen Wei[2]

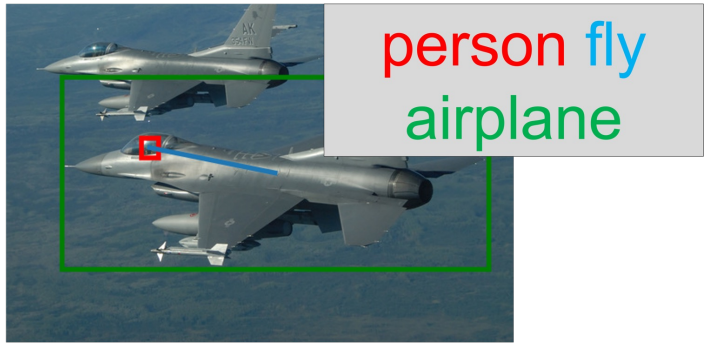
[1]Tongji University [2]MEGVII Technology [3]Shanghai Jiao Tong University




WED-PM-278

# 0 Quick preview

- Target: the challenging interaction classification sub-task in HOI detection.


- Idea: replacing the static class weight with adaptive category query learned through image-level classification.

$$cls\_score = feature * weight$$

$$\Downarrow$$

$$cls\_score = feature * category\_query$$

- Experiments: universal improvement over various methods.

# 1 HOI detection

- Two sub-tasks:

**Human**-**Object** Pair Detection

**Interaction** Classification

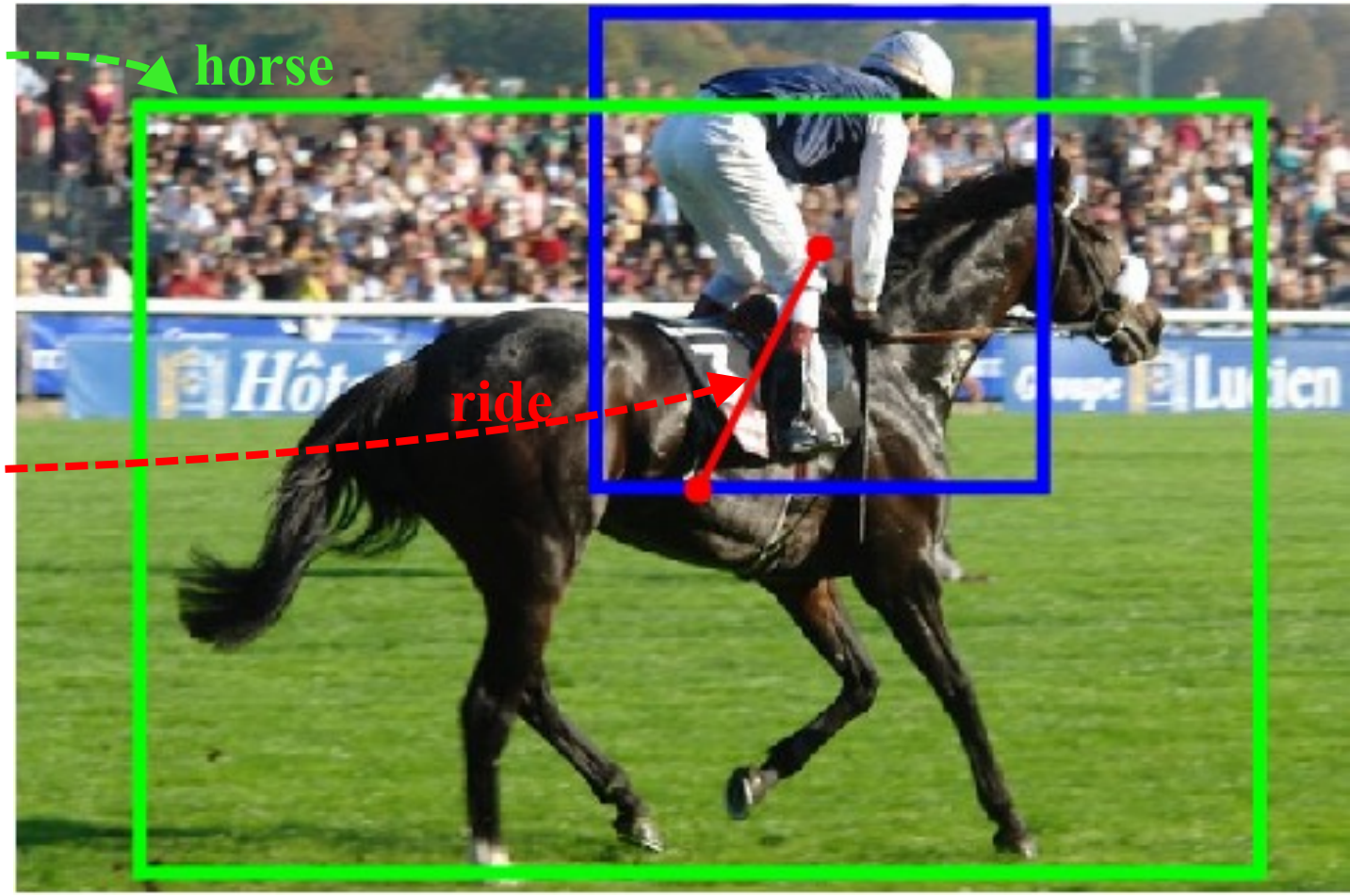
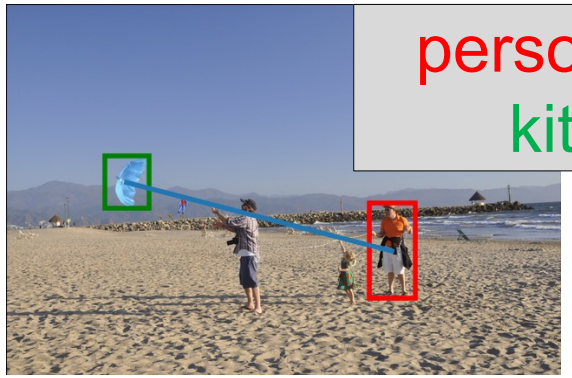


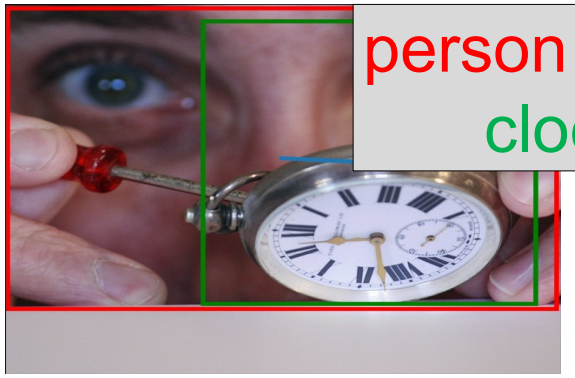

Figure from http://www-personal.umich.edu/~ywchao/hico/

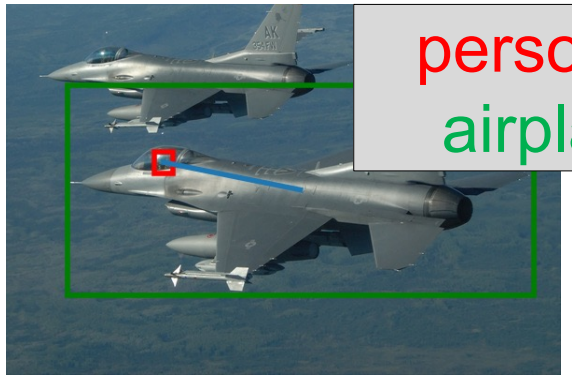
# 1 The challenging interaction classification

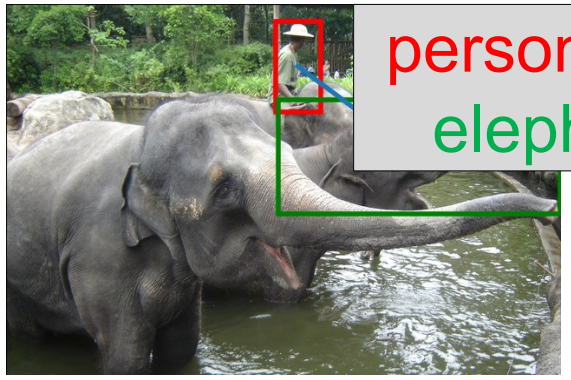
- Abstract semantics and diverse appearance, even polysemic
- Relevant objects and scene backgrounds





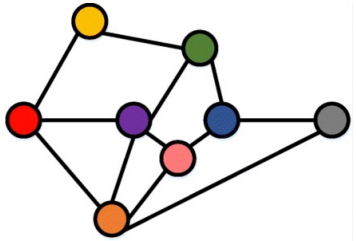


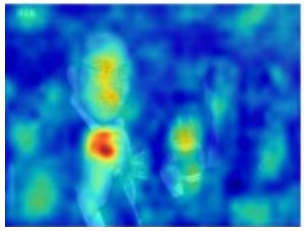




# 2 Previous interaction classification

- For $cls\_score = feature * weight$, improve feature with:
  - Context modeling

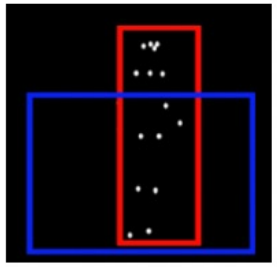


GNN



attention module

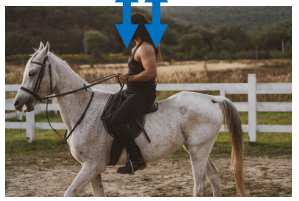
  - Introducing other cues



pose

segmentation mask

*"A man ride a horse."*



vision-language

- Better interaction representation in different detection frameworks

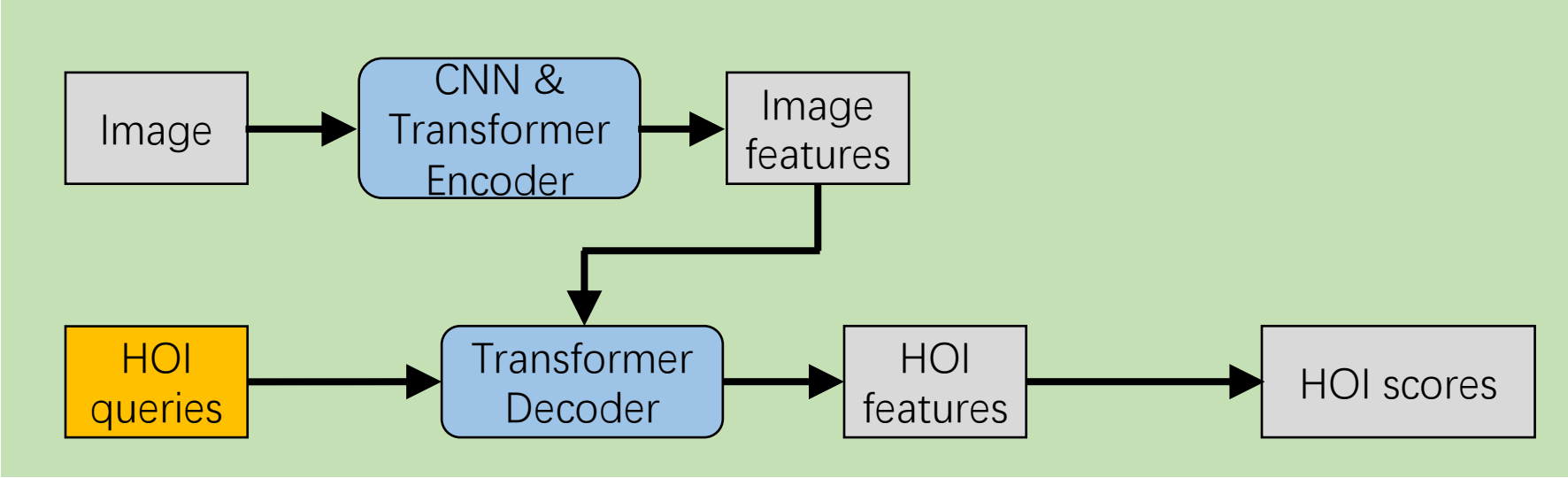
Human & object box
in two-stage methods

Union-box or middle point
in one-stage methods

Query in
transformer-based methods

# 2 Instance queries in HOI transformers

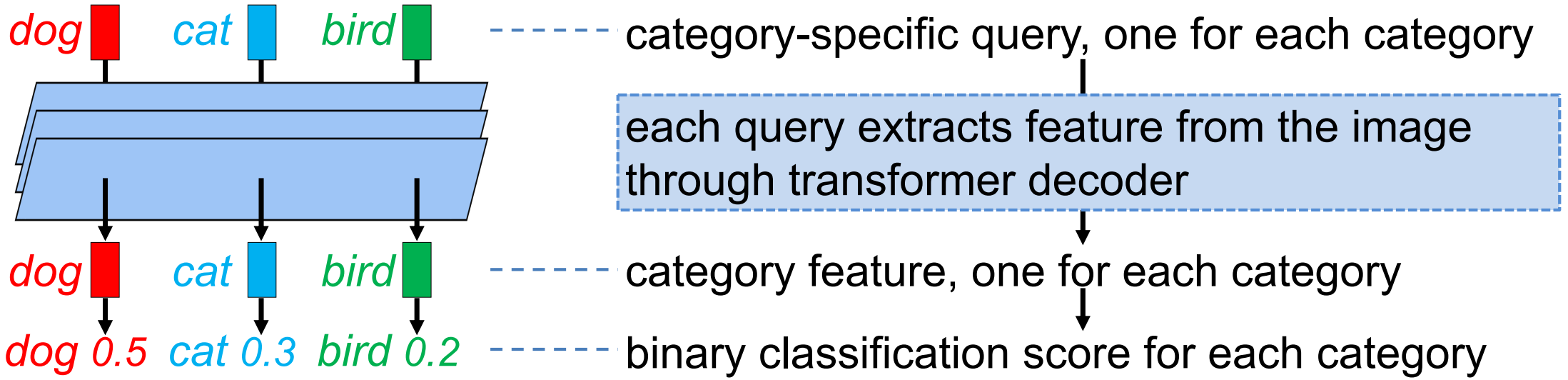
- QPIC/HOIT/CDN/……
  - Instance-wise, not explicitly bound to a category




Tamura, Masato, Hiroki Ohashi, and Tomoaki Yoshinaga. "Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
Carion, Nicolas, et al. "End-to-end object detection with transformers." European Conference on Computer Vision. Springer, Cham, 2020.
Zhang, Aixi, et al. "Mining the benefits of two-stage and one-stage hoi detection." Advances in Neural Information Processing Systems 34 (2021): 17209-17220.

# 2 Query as category representation

- Category-specific query for multi-label classification
  - Image-wise, explicitly bound to a category

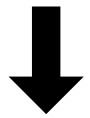



Liu, Shilong, et al. "Query2label: A simple transformer way to multi-label classification." *arXiv preprint arXiv:2107.10834* (2021).

# 3 Category query for interaction classification

- We replace the classification weight

$$cls\_score = feature * weight$$

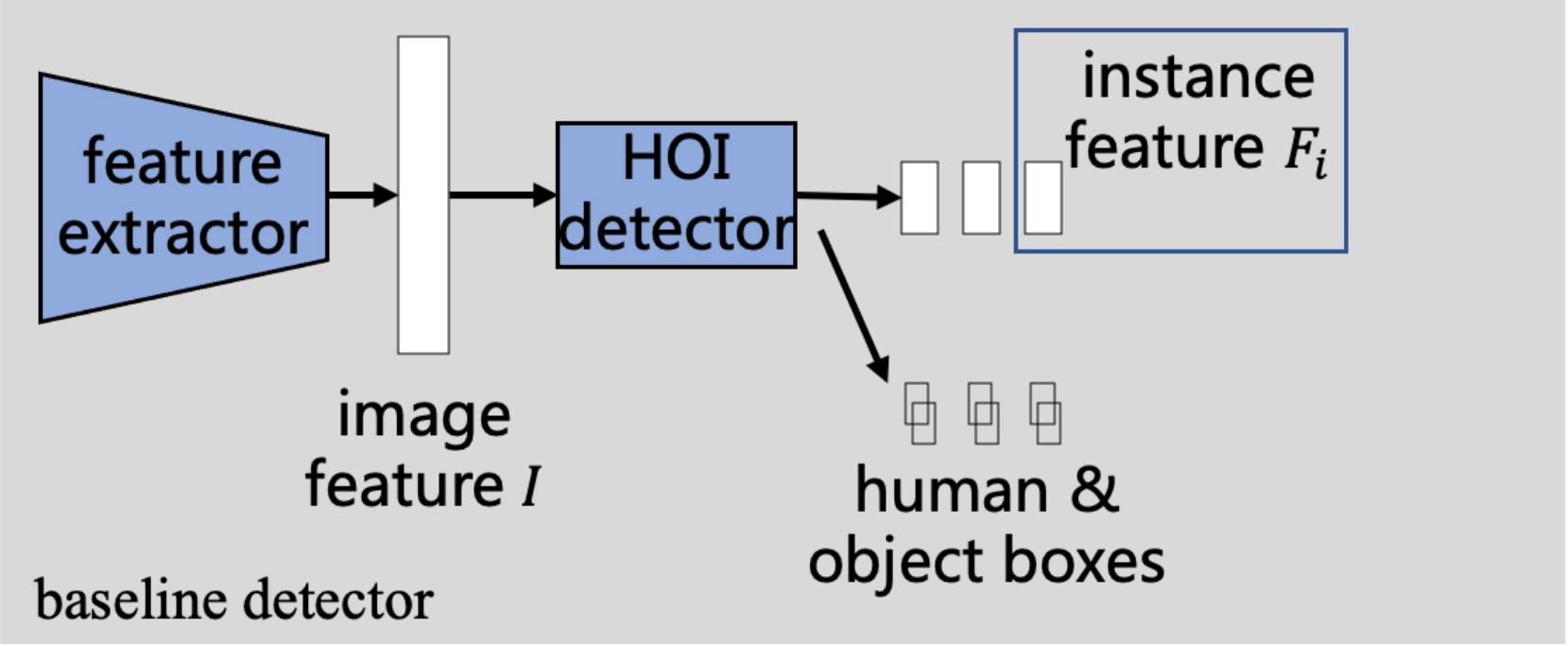


$$cls\_score = feature * category\_query$$

- Nothing else changed
  - Orthogonal to previous methods (improving interaction feature)

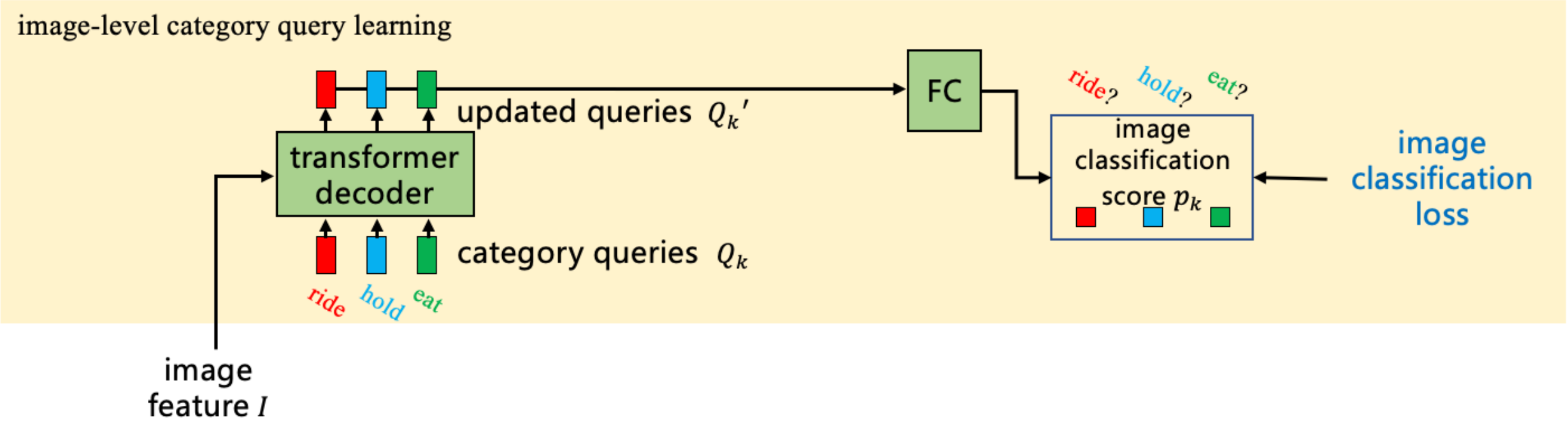
# 3 Our baselines

- Any existing HOI detection methods with:
  - Image feature $I$
  - Interaction (instance) feature $F_i$



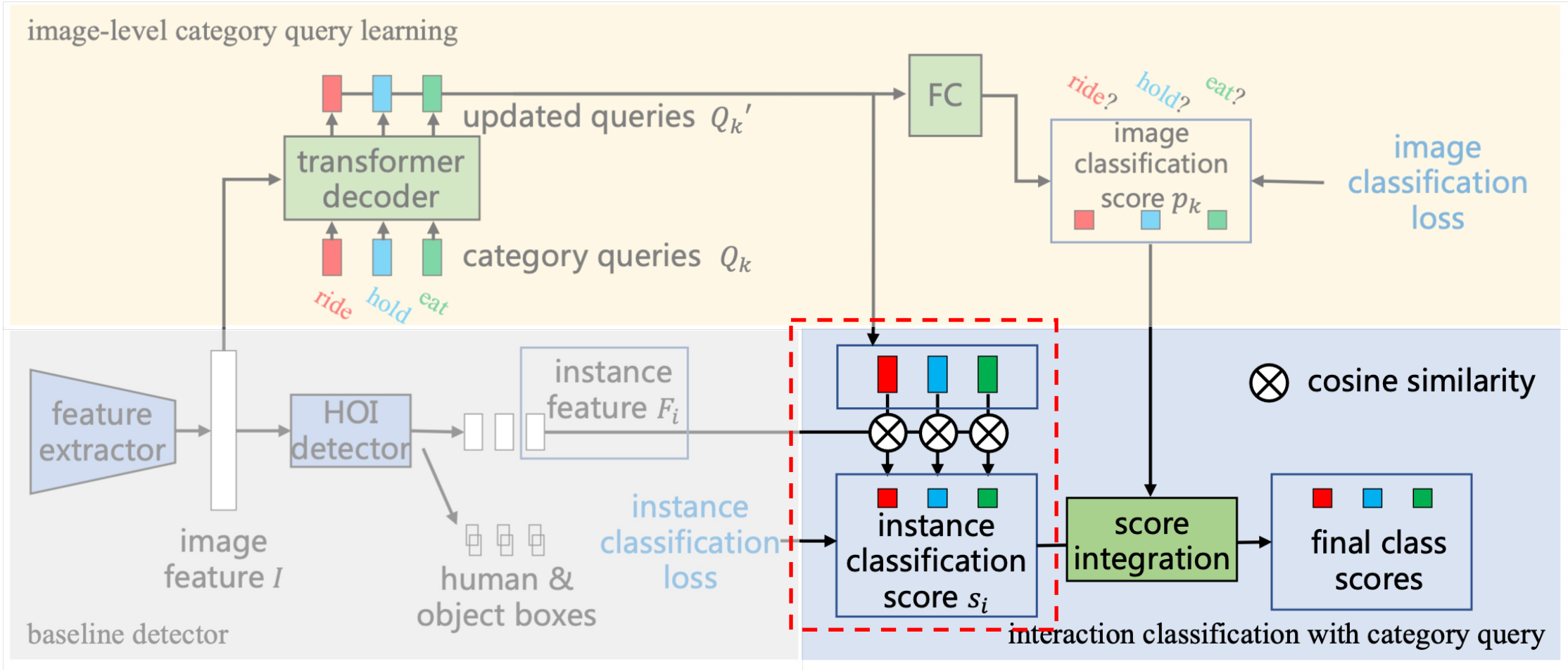
# 3 Our methods (1/2)

- Image-level category query learning
  - Auxiliary image classification task



# 3 Our methods (2/2)

- Interaction classification with category query



# 4 Improvements over 3 baselines

- Our method improves 3 baselines over 2 benchmarks

| Method | Pipeline | E2E | HICO-DET | | | V-COCO | | Efficiency | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Full | Rare | Non-Rare | S1 | S2 | #Params | FPS |
| QPIC [33] | transformer | ✓ | 28.93 | 21.62 | 31.12 | 61.39 | 63.65 | 41M | 19.5 |
| + *Ours* | transformer | ✓ | 31.08(+2.15) | 23.90 | 33.22 | 63.67(+2.28) | 65.49 | 46M(+5M) | 18.3(-6.2%) |
| SCG [42] | two-stage | ✗ | 31.28 | 24.16 | 33.40 | 56.93 | 62.51 | 57M | 4.5 |
| + *Ours* | two-stage | ✗ | 32.74(+1.46) | 26.25 | 34.68 | 59.14(+2.21) | 65.61 | 64M(+7M) | 4.1(-8.9%) |
| GEN-VLKT [22] | transformer | ✗ | 33.69 | 29.94 | 34.81 | 64.89 | 66.74 | 42M | 21.7 |
| + *Ours* | transformer | ✗ | 35.36(+1.67) | 32.97 | 36.07 | 66.40(+1.51) | 69.17 | 47M(+5M) | 20.6(-5.1%) |

transformer-based, SOTA

SOTA for two-stage

first transformer-based, end-to-end

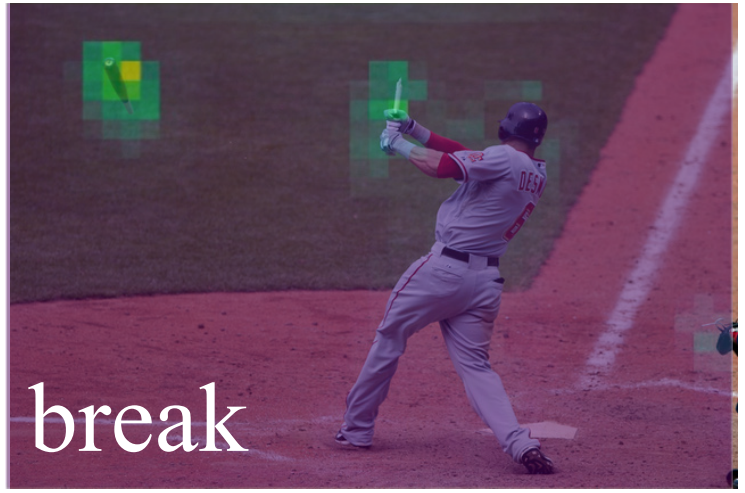
# 4 Ablation on components

- C1: image classification task
- C2: interaction classification with category query
- C3: a score integration technique

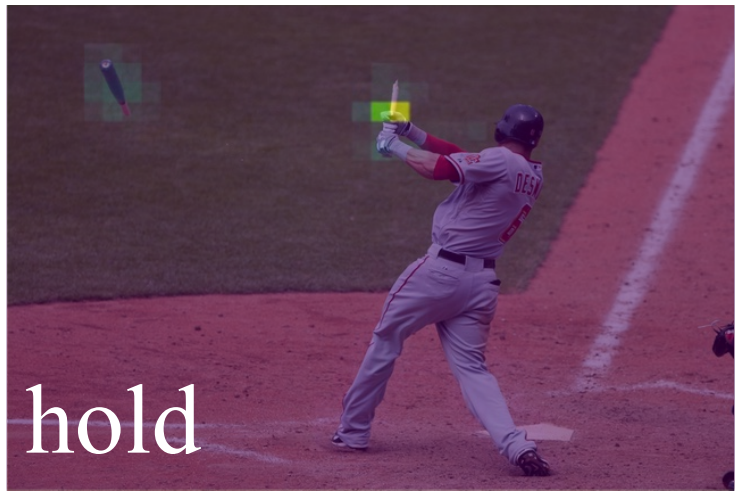
| | C1 | C2 | C3 | Full | Rare | Non-Rare |
|---|---|---|---|---|---|---|
| a | - | - | - | 33.69 | 29.94 | 34.81 |
| b | ✓ | - | - | 33.86 (+0.17) | 31.12 | 34.68 |
| c | ✓ | ✓ | - | 34.98 (+1.29) | 31.73 | 35.95 |
| d | ✓ | ✓ | ✓ | **35.36** (+1.67) | **32.97** | **36.07** |

# 4 Category query attention

- Each query learns the semantics of its category.

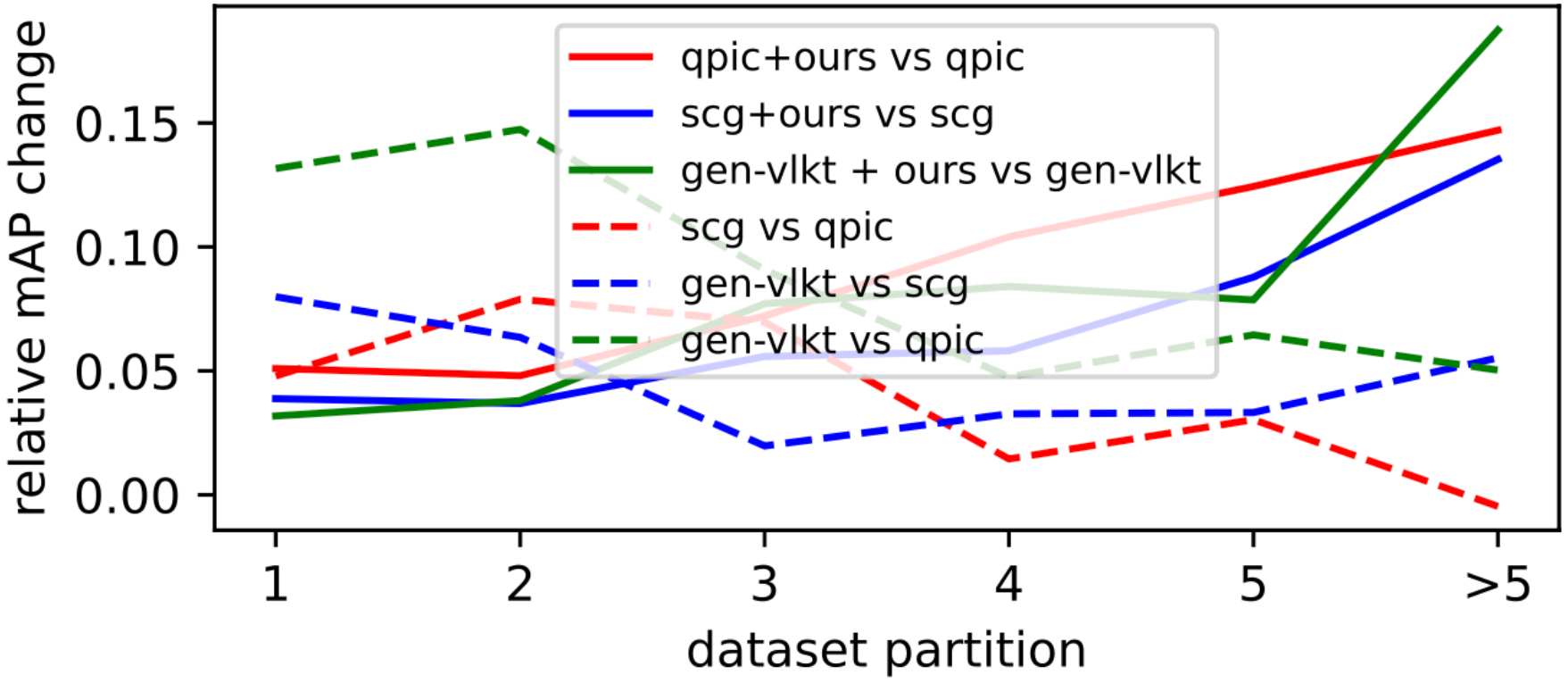




- Each query is adaptive to different images.




# 4 How the performance is improved?

- Assumption: a global modeling of all instances from one category
- Evidence: more useful on the cases with multiple instances from one category



# Thanks for listening.

Category Query Learning for
Human-Object Interaction Classification